Automated detection of white matter hyperintensities of all sizes in cerebral small vessel disease

Mohsen Ghafoorian[a)]
Diagnostic Image Analysis Group, Department of Radiology and Nuclear Medicine, Radboud University Medical Center, Nijmegen 6525, The Netherlands and Institute for Computing and Information Sciences, Radboud University, Nijmegen 6525 GA, The Netherlands

Nico Karssemeijer
Diagnostic Image Analysis Group, Department of Radiology and Nuclear Medicine, Radboud University Medical Center, Nijmegen 6525, The Netherlands

Inge W. M. van Uden and Frank-Erik de Leeuw
Donders Institute for Brain, Cognition and Behaviour, Department of Neurology, Radboud University Medical Center, Nijmegen 6525 EN, The Netherlands

Tom Heskes and Elena Marchiori
Institute for Computing and Information Sciences, Radboud University, Nijmegen 6525 EC, The Netherlands

Bram Platel
Diagnostic Image Analysis Group, Department of Radiology and Nuclear Medicine, Radboud University Medical Center, Nijmegen 6525, The Netherlands

(Received 8 April 2016; revised 5 August 2016; accepted for publication 11 October 2016; published 2 November 2016)

Purpose: White matter hyperintensities (WMH) are seen on FLAIR-MRI in several neurological disorders, including multiple sclerosis, dementia, Parkinsonism, stroke and cerebral small vessel disease (SVD). WMHs are often used as biomarkers for prognosis or disease progression in these diseases, and additionally longitudinal quantification of WMHs is used to evaluate therapeutic strategies. Human readers show considerable disagreement and inconsistency on detection of small lesions. A multitude of automated detection algorithms for WMHs exists, but since most of the current automated approaches are tuned to optimize segmentation performance according to Jaccard or Dice scores, smaller WMHs often go undetected in these approaches. In this paper, the authors propose a method to accurately detect all WMHs, large as well as small.
Methods: A two-stage learning approach was used to discriminate WMHs from normal brain tissue. Since small and larger WMHs have quite a different appearance, the authors have trained two probabilistic classifiers: one for the small WMHs (⩽3 mm effective diameter) and one for the larger WMHs (>3 mm in-plane effective diameter). For each size-specific classifier, an Adaboost is trained for five iterations, with random forests as the basic classifier. The feature sets consist of 22 features including intensities, location information, blob detectors, and second order derivatives. The outcomes of the two first-stage classifiers were combined into a single WMH likelihood by a second-stage classifier. Their method was trained and evaluated on a dataset with MRI scans of 362 SVD patients (312 subjects for training and validation annotated by one and 50 for testing annotated by two trained raters). To analyze performance on the separate test set, the authors performed a free-response receiving operating characteristic (FROC) analysis, instead of using segmentation based methods that tend to ignore the contribution of small WMHs.
Results: Experimental results based on FROC analysis demonstrated a close performance of the proposed computer aided detection (CAD) system to human readers. While an independent reader had 0.78 sensitivity with 28 false positives per volume on average, their proposed CAD system reaches a sensitivity of 0.73 with the same number of false positives.
Conclusions: The authors have developed a CAD system with all its ingredients being optimized for a better detection of WMHs of all size, which shows performance close to an independent reader.
 [http://dx.doi.org/10.1118/1.4966029]

Key words: automated detection, white matter hyperintensity, computer aided detection, Small vessel disease, adaboost

1. INTRODUCTION

Cerebral small vessel disease (SVD) is a frequently found neurological disorder in elderly people, which makes it a growing concern for countries with aging populations. As measured in the Rotterdam study[1] on a population of 1077 randomly selected elderly people, the prevalence of SVD has been reported to reach up to 95%. The SVD spectrum includes

amongst others, white matter hyperintensities (WMH) (also known as white matter lesions or leukoaraiosis), lacunes of presumed vascular origin (lacunes), cerebral microbleeds, and brain subcortical atrophy.[2] There is evidence for increased risk of cognitive, motor, and mood disturbances, ultimately leading to dementia and Parkinsonism in a small number of patients diagnosed with SVD.[3–7] Considering these, some studies are investigating the effect of SVD on the transition from nondemented elderly people with SVD toward the mentioned disorders.[8,9] One of the most important and common findings in MRI images of SVD patients is WMH.[9] WMHs are areas of demyelinated cells found in the white matter (WM) of the brain which appear as high value signals on *T*2 weighted or fluid-attenuated inversion recovery (FLAIR) MR images.

WMHs are not only found in SVD patients but are also common findings on brain MR images of the patients diagnosed with multiple sclerosis (MS),[10] Alzheimer's disease,[11] other forms of dementia,[12] stroke[13] and Parkinsonism.[14] In many studies a relationship between WMH severity and neurological symptoms, including cognitive decline,[4] gait dysfunction,[15] as well as depression and mood disturbances[5,16] were reported.

WMHs are often used as biomarkers for prognosis and disease progression in white matter disorders and additionally longitudinal quantification of WMHs is used to evaluate therapeutic strategies. For this reason accurate quantification of WMHs in terms of total load (total volume of WMHs), number of lesions and location distribution is interesting, not only for research purposes but also for the development of clinical applications. Manual segmentation of WMHs is a potential solution, but has several drawbacks: it is very time consuming, as it can take up to 5 h, according to our local domain experts. It is also subjective and prone to miss small WMHs. For instance referring to Fig. 1 the readers miss or disagree on 30% of WMHs with in-plane effective diameter of 3 mm or less. Therefore automated quantification of WMHs is an attractive topic for research and hence many automated methods have been proposed over the years. A number of methods use unsupervised approaches to cluster WMHs as outliers,[17–26] while other methods segment WMHs using supervised machine learning techniques.[27–35,54,55] Although a multitude of approaches has been suggested for this problem, a truly reliable fully automated method that performs as good as human readers has not been identified.[36,37]

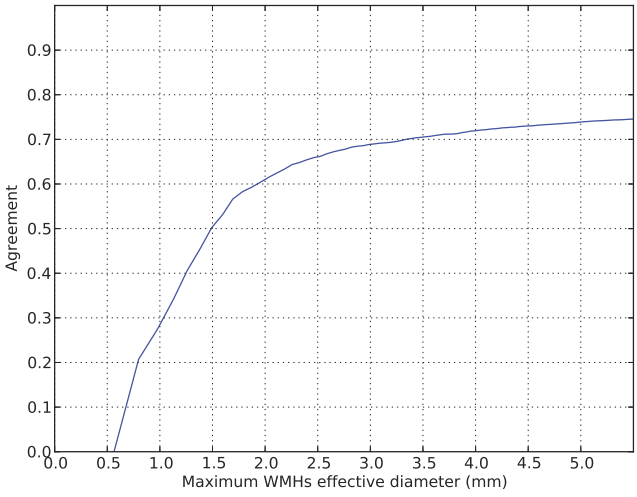


FIG. 1. Inter-reader agreement based on maximum WMHs in-plane effective diameter. The agreement factor represents the proportion of the total WMHs in both readers' annotations, smaller than a specified size, that are intersecting with an annotation of the other reader by at least one voxel.

The assumptions that motivate the use of human readers annotations (although they are not perfect referring to Fig. 1) as the ground truth for training and evaluation are the following: First, there are no alternatives yet proven to provide better segmentation than human readers annotation. Second, the readers are assumed not to persistently make errors for a specified class of WMHs (e.g., always overlooking small lesions). Referring to Fig. 1 for the small WMH category, the readers still agree on majority (70%) of cases. This lets the machine learning systems to be able to statistically learn about the categories that were occasionally wrongly labeled.

Nearly all of the existing methods, of which some are referenced above, are developed to segment WMHs and are tuned to maximize overlap between areas of WMH as measured by the Jaccard or Dice coefficient.[36,37] As a result, small WMHs might be ignored since they hardly contribute to the Jaccard or Dice performance[37] as they form a small part of WMH volume.

Especially for SVD, small WMHs are abundant and appear to be important. Analyzing the annotations made by human readers in our dataset of over 500 SVD patients, the in-plane effective diameter of over 60% of WMHs is equal to or less than 3 mm, where the in-plane effective diameter is the diameter of a circle with the same area. This large amount of small WMHs only contributes to 15% of the total volume. This implies that with a more accurate detection of small WMHs, it is possible to better assess the location and number of WMHs. Moreover, small WMH detection is vital for tracking lesion growth and general measurement of WMH progress speed. The detection of small WMHs can be indicative for neurological deficits that will emerge over time. As Schmidt *et al.*[38] suggest, progression of WMH as shown by MRI may provide a surrogate marker in clinical trials on cerebral small vessel disease in which the currently used primary outcomes are cognitive impairment and dementia.

Considering the above, accurate detection of WMHs, both small and large, is an interesting subject for research, e.g., to be able to longitudinally monitor WMH progression. Further research needs to be done to investigate the clinical importance of small WMHs. Empirical results of our Radboud University Nijmegen Diffusion tensor and Magnetic resonance imaging Cohort (RUN DMC) study show that some small lesions grow in size over time, which could indicate the relevance of small lesions for the prediction of disease progression. It is recommended to detect WMHs of all size and to consider the number of WMHs, together with their volume and location distribution as the measures describing WMH characteristics and severity, as described in the SVD standards for neuroimaging research.[2] It should be noted that the number of detected WMHs would be highly influenced by the quality of the detection for small WMHs as they form the majority of WMHs in counts (see Fig. 3).

There are some fundamental differences in the characteristics of small (⩽3 mm in-plane effective diameter) and large (>3 mm in-plane effective diameter) WMHs. First of all, small WMHs usually appear to have a different intensity range likely because of the partial volume effect[39] (which occurs when voxels cover tissue boundaries and therefore represent a mixture of tissues). Secondly, small WMHs usually appear as blob like structures, while larger WMHs can show up in more arbitrary shapes. Thirdly, small lesions tend to appear at different locations than larger WMHs, which occur more often along the ventricles.[4] The heterogeneity of the smaller and larger lesions makes their representation scattered over different regions in the feature space resulting in a highly nonlinear problem and therefore making it more difficult to solve.[40] Given this, we were motivated to reduce the complexity of the problem by dividing the WMHs into small and large WMH categories and learn each concept separately by means of supervised machine learning.

As discussed before, measures regarding the overlapping area, such as Dice or Jaccard, do not sufficiently reflect the detection of smaller lesions. Therefore we utilize a free-response receiving operating characteristic (FROC) analysis[41] to evaluate the performance of the proposed method.

In this paper we present a method for the accurate automatic detection of WMHs in SVD. Where the state-of-the-art approaches do not specifically focus on the small WMHs, we use a novel approach in which we detect WMHs by combining the output of two separate classifiers, one for large and one for small WMHs. To describe each of the lesion types we introduce a set of specialized features. The results of our method are compared to manual annotations of two human readers, showing a close performance of the resulting computer aided detection (CAD) system to human readers.

2. MATERIALS AND METHODS

The overall pipeline for this automated detection task consists of data acquisition, image preprocessing, feature calculation, training, and evaluation. Figure 2 shows an overview of the whole pipeline. Method components will be expanded in separate subsections subsequently.

2.A. Data

The research presented in this paper uses data from a follow-up study called RUN DMC.[9] Baseline scanning was performed in 2006. The patients were rescanned in 2011/2012 and 2015. This study was approved by the Medical Review Ethics Committee region Arnhem-Nijmegen. All participants gave written informed consent prior to inclusion.

2.A.1. Subjects

Subjects for the RUN DMC study were selected at baseline based on the following inclusion criteria:[9] (a) aged between 50 and 85 yrs and (b) cerebral SVD on neuroimaging (appearance of WMHs and/or lacunes).

Exclusion criteria comprised of the following: presence of (a) dementia, (b) parkinson(-ism), (c) intracranial hemorrhage, (d) life expectancy less than six months, (e) intracranial space occupying lesion, (f) (psychiatric) disease interfering with cognitive testing or follow-up, (g) recent or current use

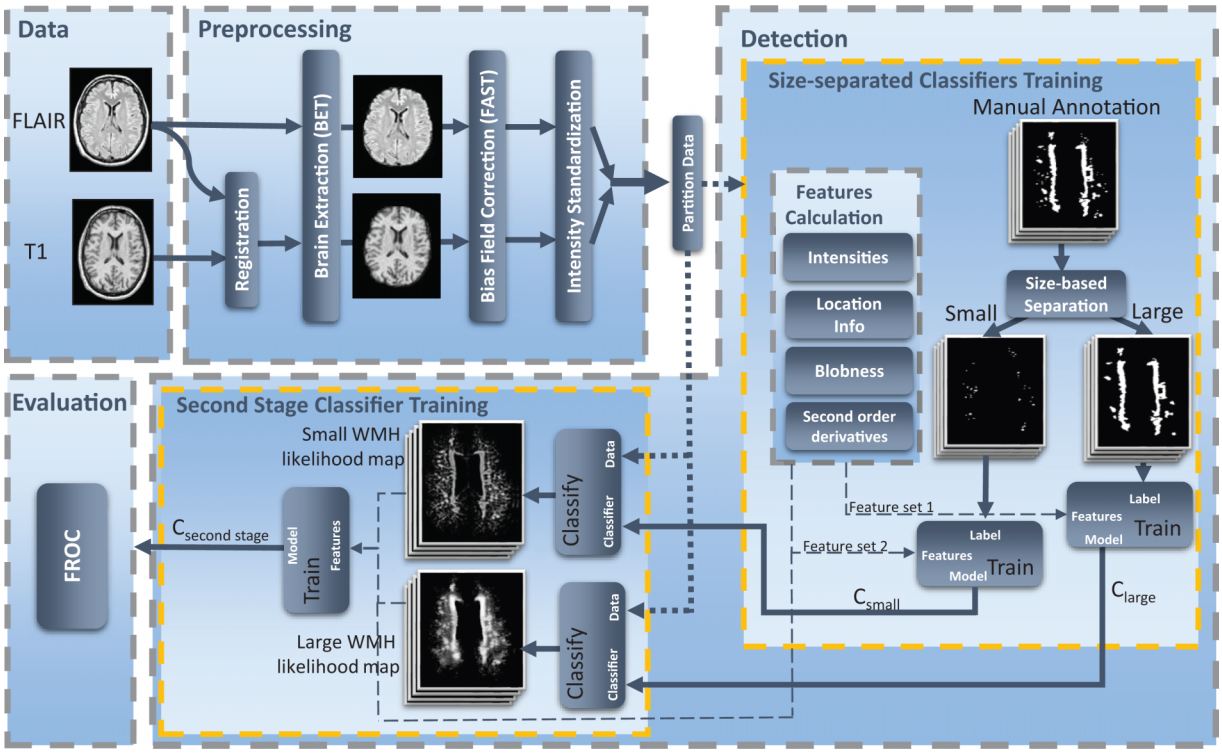


FIG. 2. An overview of the steps taken for the overall image analysis task.

of acetylcholine-esterase inhibitors, neuroleptic agents, L-dopa or dopa-a(nta)gonists, (h) non-SVD related WMH (e.g., MS), (i) prominent visual or hearing impairment (j) language barrier, and (k) MRI contraindications. Based on these criteria, MRI scans of 503 patients were taken. All of the subjects showed (at least mild) appearances of WMH in their MR images. The distribution of the Fazekas scores[42] of the scanned subjects were as follows: 66% Fazekas 0 or 1 (mild lesion load), 21% with Fazekas 2 (moderate load), and 13% with Fazekas 3 (severe lesion load).

2.A.2. Magnetic resonance imaging

The machine used for the baseline was a single 1.5 T scanner (Magnetom Sonata, Siemens Medical Solution, Erlangen, Germany). The protocol included a 3D $T1$ magnetization-prepared rapid gradient-echo sequence (TR/TE/TI 2250/3.68/850 ms; flip angle 15; voxel size $1.0\times1.0\times1.0$ mm) and FLAIR pulse sequences (TR/TE/TI 9000/84/2200 ms; voxel size $1.2\times1.0\times5.0$ mm, interslice gap 1 mm). All the scans were acquired with the same acquisition settings and scanner with no major software and hardware upgrades.

2.A.3. Reference annotations

Reference annotations were manually created in a slice by slice manner by two trained readers using a digital pen. The training procedure was as follows: The readers were instructed on the manual annotation of WMHs and the use of the provided annotation tools. Following the definition in Ref. 9, WMHs were defined as hyperintense lesions on FLAIR MRI that did not show corresponding cerebrospinal fluid (CSF) like hypointense lesions on the $T1$ weighted image, excluding Gliosis surrounding lacunes and territorial infarcts. After these instructions both readers annotated a training set of 50 unannotated cases, and each reader was blinded to the annotations of the other. To further reduce the inter-rater variability, these annotations were discussed together with an experienced neurologist in a follow-up meeting. After this training, 453 cases were annotated by either one of the readers (reader 1), and 50 cases were annotated by both.

An investigation on the number of WMH annotations on different patients for reader 1 shows that on average 123 WMHs were annotated (lesions were counted on every slice) with a standard deviation of 75. The average and standard deviation were 100 and 65 for reader 2 respectively. Figure 3 shows a histogram for the distribution of the in-plane effective diameters of WMH annotations created by reader 1 and compares it to a similar histogram for MS lesions calculated from a publicly available dataset (ISBI 2015 longitudinal MS lesion segmentation challenge). This figure illustrates that SVD has a higher concentration of small lesions compared to MS.

2.B. Preprocessing

Due to possible patient movements between scans of different imaging modalities and uneven intensity profiles intra and intersubjects, image preprocessing is a crucial part of our algorithm. Below we give a short description of the steps taken to prepare the images for feature calculation.

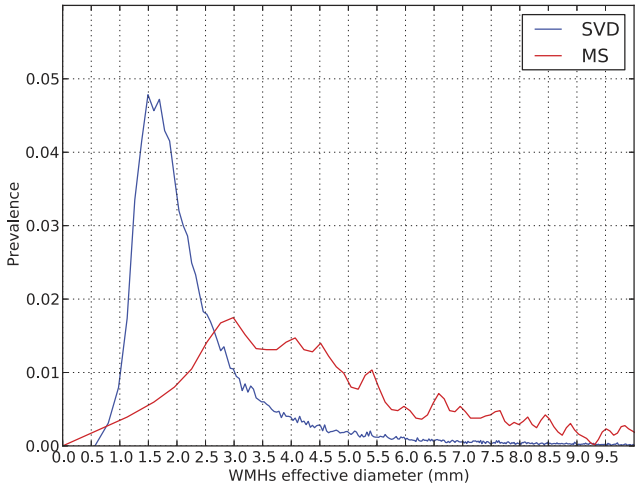


FIG. 3. Distribution of WMH sizes in the reference annotation in two datasets of SVD and MS.

2.B.1. Registration, skull removal, and bias field correction

First of all, establishing a voxel classification method that uses intensity features requires locational alignment between each voxel in one modality and the corresponding voxel in other modalities. Possible patient movements between different scans make this a nontrivial step.

To tackle this, for each subject, $T1$ images were rigidly registered to the FLAIR images by optimizing mutual information with trilinear interpolation resampling, as implemented in FSL-FLIRT.[43] We avoid transforming the FLAIR image to $T1$ in order to prevent possible artifacts on FLAIR and the annotations that are made on the FLAIR image. In addition, all subjects were registered to the ICBM152 atlas[44] to acquire a mapping from each subject space to the atlas space.

Once images were registered, skull, eyes, and other nonbrain tissues were removed. For this, we made use of FSL-BET (Ref. 45) on the patient's $T1$ image and then applied the resulting mask to the other modality. For FSL-BET, we used the robust brain center estimation option that iteratively calls BET with the initial center of brain set each time to the center-of-gravity of the previously estimated brain extraction. We chose $T1$ since it has the highest resolution among the three modalities.

Bias field correction is another necessary step due to magnetic field inhomogeneity. To this end, we applied FSL-FAST (Ref. 46) which uses a hidden Markov random field and an associated expectation–maximization algorithm, solely for bias-field correction purpose. FSL-FAST was executed with two modalities (FLAIR and $T1$) as its input channels, modeling the brain with three tissue classes.

2.B.2. Intensity standardization

In addition to intensity inhomogeneities caused by the MR bias field, it is very common to see intensity inhomogeneity

between different subjects. Correction of these intersubject intensity inhomogeneities is essential since MRI intensity is an important feature.

The general approach that we followed, similar to most existing methods, was to pick a reference image and transform other images, so that all intensity profiles resemble each other. In order to get a finer intensity transformation, we considered three different transformations for the three brain tissue types: gray matter (GM), WM, and CSF.

First, we extract the three tissues of the reference image using bivariate Gaussian mixture modeling[47] of the two variables $T1$ and FLAIR intensities. We then project each 2D Gaussian on the dimension corresponding to FLAIR intensity, to obtain three 1D Gaussians for the reference subject, with means and standard deviations ($\mu_{\text{ref,gm}}$, $\sigma_{\text{ref,gm}}$), ($\mu_{\text{ref,wm}}$, $\sigma_{\text{ref,wm}}$), and ($\mu_{\text{ref,csf}}$, $\sigma_{\text{ref,csf}}$). With a similar approach, we obtain Gaussians for each template image ($\mu_{\text{temp,gm}}$, $\sigma_{\text{temp,gm}}$), ($\mu_{\text{temp,wm}}$, $\sigma_{\text{temp,wm}}$), and ($\mu_{\text{temp,csf}}$, $\sigma_{\text{temp,csf}}$). Then for a given intensity x, the transformed intensity depends on the assumption made for the tissue it belongs to, using the following equation:

$$T_k(x) = \frac{(x - \mu_{\text{temp},k})}{\sigma_{\text{temp},k}} \times \sigma_{\text{ref},k} + \mu_{\text{ref},k}, \quad (1)$$

where $k \in \{\text{WM, GM, CSF}\}$. Gaussian mixture modeling provides the posterior probabilities of intensities belonging to each tissue. Hence the following equation was used to acquire the transformed intensity value:

$$T(x) = \sum_{k \in \{\text{WM,GM,CSF}\}} T_k(x) \times p(x \in k). \quad (2)$$

The same procedure was applied to standardize the $T1$ images.

2.B.3. Selection of training and test subjects

To enable comparison of our method with human readers, we use the 50 subjects with two annotations for testing purposes and the rest for training our model. However, a number of cases contained artifacts that were obscuring fine structures of the brain. We opted not to include these cases in our training set. We visually filtered out cases that showed scanning artifacts due to head movements during the scanning as well as the cases for which one of the preprocessing steps failed (most often registration, or brain extraction failure). After this selection, 312 scans remained to train the system. From the 50 double annotated cases that were used for testing performance, 32 were found not to contain severe artifacts, the remaining 18 more challenging cases were not removed from the test set, but were evaluated separately. Table I represents the number of cases filtered for each of the reasons. We should note that we also evaluate our method on the problematic cases of the test set to show to what extent our CAD system is usable for these cases.

TABLE I. Case removal cause distribution in the training and test sets.

Set	Movement artifacts	Brain extraction failure	Registration failure
Train	104	36	1
Test	16	2	0
Total	120	38	1

2.C. Detection

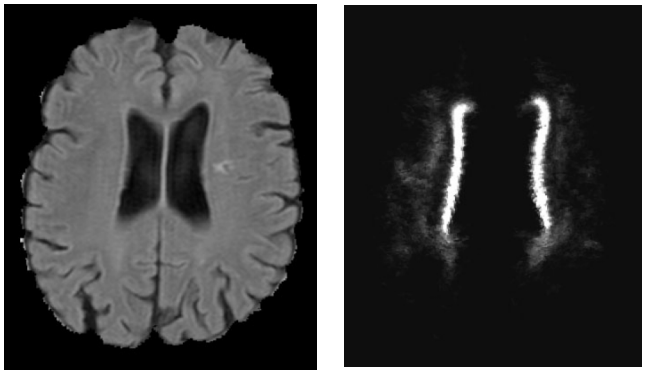
As Fig. 3 suggests, the majority of WMHs in SVD is tiny. Due to the different location and appearance of small and larger WMHs, intuitively they require a different set of features to describe their appearances. Considering this, a single WMH classifier potentially misses small WMHs. We therefore specify two different classifiers, which were trained on the same set of subjects, but using different sets of features for small (⩽3 mm in-plane effective diameter) and larger WMHs. The final goal is an algorithm that specifies for each voxel the likelihood that it belongs to a WMH, independent of whether it belongs to a small or a large WMH. We have built two first-stage classifiers that each provide us likelihoods for small[48] and larger WMHs and one second-stage classifier that combines the two likelihoods into a single WMH likelihood. Each learning problem is described in one of the following subsections 2.C.1 and 2.C.2. As training cases 312 subject images that were annotated by reader 1 were used and we evaluated the system on 50 double annotated subjects in total.

2.C.1. Small and large WMH detectors

2.C.1.a. Features. Using voxels as training samples, we trained two voxel-based classifiers, one for small and one for larger WMHs. Every single voxel for the larger WMH detector was characterized by eleven features. The first two features correspond to the bias field corrected, standardized FLAIR and $T1$ intensities. WMHs in SVD are not uniformly distributed over different locations. For example, WMHs often occur in the periventricular region. Furthermore, although voxels in the septum pellucidum might appear hyperintense, they do not originate from white matter demyelination and thus do not belong to WMHs.

This then motivates the following features: X, Y, and Z coordinates as measured in the reference space defined by the ICBM152 atlas, and the voxel's shortest Euclidean distance to the left and right ventricles, brain cortex, and midsagittal brain surface. In addition, from a large number of subjects with WMH annotations, we computed the distribution of WMHs over different locations. For each atlas space location, the proportion of subjects with a WMH in the corresponding position was calculated yielding a prior probability map. This WMH occurrence prior probability map, visualized for a sample case in Fig. 4, provides another feature. The full list of features used is shown in Table II.

For the small WMH detector, we take the same eleven features as for the larger WMH detector, plus a set of additional features considered exclusively for characterizing small WMHs. Because small WMHs usually appear as a blob-like structure, we include as features various measures of blobness at different scales: Laplacian of Gaussian, determinant of the Hessian matrix, and the output of a multiscale grayscale annular filter,[49] each at three different scales: t = 1, 2, and 4 mm. In addition, because WMHs occur in WM by definition,



(a) (b)

FIG. 4. A sample subject prior probability for occurrence of WMH. (a) One FLAIR slice of a sample patient and (b) Corresponding WMH prior probability in the patient space.

the segmentation results obtained from the standardization step provide a discrete variable taking three values.

In some cases, GM parts of cortex appear as isolated structures inside the WM, due to the 3D folding pattern and the sliced based imaging. Since GM has higher signal intensity in FLAIR compared to normal WM, it is important to distinguish these GM parts from true WM to prevent false detections. These GM structures usually appear in an elongated shape. Therefore, we include two features for characterizing these vessel-like structure: vesselness ($\sigma = 1$) and gauge derivative in the direction of the normal vector.[50]

2.C.1.b. Sampling. Following are the details of how we select for each classifier the samples that represent the tissue of interest to be detected (positive samples) and the samples that represent the background tissue (negative samples). For both larger and small WMH detectors we utilized 75% of training subjects. In our voxel-based classification scheme, we only select voxels from these subjects for training. WMHs were separated into small and larger WMH categories using a size threshold on the manual annotations: a WMH with an effective diameter smaller than or equal to 3 mm is considered small and hence a positive sample for the small WMH detector. WMHs with an in-plane effective diameter larger than 3 mm were considered large and hence seen as a positive sample for the large WMH detector. We picked this threshold referring to WMH size distribution illustrated in Fig. 2, where 3 mm is two times larger than the small WMH distribution peak at 1.5 mm effective diameter. Normal brain voxels are potential negative samples for both size-separated classifiers.

To prevent trivial negative samples, we removed all voxels with FLAIR signal intensity lower than a threshold, as well as the voxels that belong to ventricles. This threshold was selected based on intensity distribution of lesions after the intensity standardization, to make sure that all lesions in our dataset are preserved in the remaining voxels. Because there are many more negative samples compared to positives, we included all positive samples of the subject considered for training into the training set and randomly picked 2% of the remaining negative samples.

TABLE II. Features used for small WMH, large WMH and second stage classifiers.

Feature group	Feature	Small WMH detector	Large WMH detector	Second stage classifier
Intensities	FLAIR intensity	Yes	Yes	Yes
	*T*1 intensity	Yes	Yes	Yes
Location	*X* in atlas space	Yes	Yes	Yes
	Y in atlas space	Yes	Yes	Yes
	Z in atlas space	Yes	Yes	Yes
	Shortest Euclidean distance to the brain cortex	Yes	Yes	Yes
	Shortest Euclidean distance to the right ventricle	Yes	Yes	Yes
	Shortest Euclidean distance to the left ventricle	Yes	Yes	Yes
	Shortest Euclidean distance to the midsagittal brain surface	Yes	Yes	Yes
	Prior probability based on atlas location	Yes	Yes	Yes
Blobness	Laplacian of Gaussian (small scale)	Yes	No	Yes
	Laplacian of Gaussian (medium scale)	Yes	No	Yes
	Laplacian of Gaussian (large scale)	Yes	No	Yes
	Determinant of Hessian (small scale)	Yes	No	Yes
	Determinant of Hessian (medium scale)	Yes	No	Yes
	Determinant of Hessian (large scale)	Yes	No	Yes
	Grayscale annular filter (small scale)	Yes	No	Yes
	Grayscale annular filter (medium scale)	Yes	No	Yes
	Grayscale annular filter (large scale)	Yes	No	Yes
Second orders	Vesselness	Yes	No	Yes
	Gauge derivative in the direction of the normal vector	Yes	No	Yes
	Tissue segmentation	Yes	No	Yes
Size-separated WMH likelihoods	Likelihood of being small WMH	No	No	Yes
	Likelihood of being large WMH	No	No	Yes

We left out the small WMH samples from the training set of the large WMH detector and vice versa. That is, they were neither considered as positive nor negative samples. The reason for this was to avoid confusing the classifier with their partial similarity. This might cause the large WMH detector to detect some small WMHs as well and vice versa, but this is no problem as the final goal is to detect all WMHs.

2.C.1.c. Training and classification. Accurate detection of small WMHs is a complex task. This is because image noise can mimic small lesions. In addition, readers are less reliable at identifying small WMHs, which leads to an inaccurate ground truth for the learning algorithm to train on.

We have chosen to use random forest[51] using the following parameter settings: maximum 20 subtrees, with $\sqrt{\#features}$ features randomly selected at each node, information gain as the tree splitting criterion, and *#features* as the maximum depth of the tree. In order to be able to concentrate more on learning the concept behind harder samples, five iterations of Adaboost[52] were run. In each iteration of Adaboost a random forest was created, which concentrates more on learning the concept via samples that were misclassified in the previous iterations. This will help the classifier to perform better at labeling harder samples.

To assess the performance of Adaboost on random forest as the classifier, we also trained on the same data a single random forest (with the same settings) as well as a Gentleboost[53] classifier using 100 regression stumps as the weak classifiers. We optimized the parameters of the methods considering a separate validation set of ten subjects. The optimization criterion was either qualitative results (e.g., for vesselness σ to check if they respond well to the objects of interest) or FROC curves for classifier parameters (e.g., number of iterations in Adaboost or max number of trees in random forest).[54]

2.C.2. Second-stage classification

After the two likelihoods computed by the small and large WMH detectors are acquired, they were subsequently merged into a single likelihood, representing the WMHs regardless of their size. Figure 5 depicts a scatter plot representing the small and large WMH likelihoods for each sample, where the positive and negative samples are distinguished with green and red colors respectively.

As a simple approach one could threshold the two likelihood maps and merge these results. This would correspond to discriminating the two classes with a pair of horizontal and vertical lines on the scatter plot in Fig. 5. It is clear, however, that this does not result in a good separation of the two classes. Instead, we consider this merging as another learning problem, which learns the WMH likelihood given the likelihoods of each voxel being in a small or large WMH.

2.C.2.a. Combination features. The likelihood of being a small WMH and the likelihood of being a large WMH were the two basic features used to represent each sample used for merging the likelihoods. As Fig. 5 shows, although these two likelihoods are good features for discrimination of WMHs and normal WM, the separation is not perfect. By adding more features we improved the performance of the classifier. For instance, if the classifier has the information that a voxel comes from a small-grained structure, it can learn that it should put more weight on the small WMH likelihood. To improve the results we included all of the features used for the detection of small WMH classifier in the second-stage classifier features set as well.

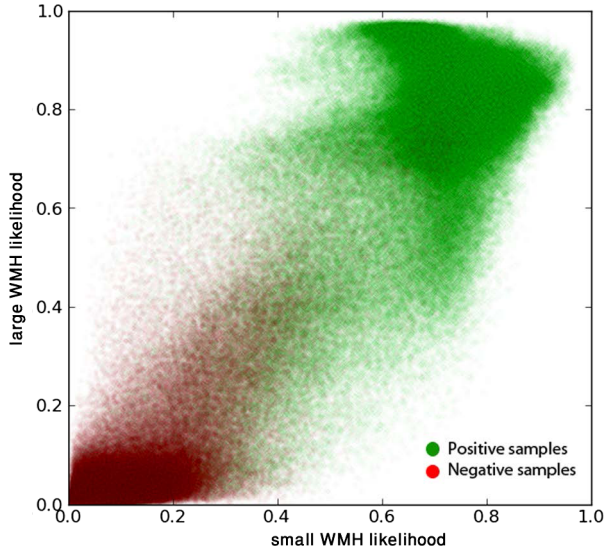


FIG. 5. 2D projection of scatter plot for the second-stage classifier samples on small and large WMH likelihoods.

2.C.2.b. Sampling. As mentioned earlier, we split the training dataset into two subsets of 75% (234 cases) and 25% (78 cases) and used the first set to train the two size-separated classifiers. We used the second subset to train the second-stage classifier. The motivation to perform this separation was to avoid potential bias due to usage of the classification likelihoods on the same training data. From the set of images considered for training of the second-stage classifier, we select all the voxels annotated as WMHs, no matter how small or large they are, as the positive samples. For the negative samples, 0.3% of the nonWMH voxels are uniformly selected at random, to create a relatively balanced dataset.

2.C.2.c. Training and classification. Adaboost was used for the second-stage classification as well, and consisted of five iterations of training random forest as the basic classifier.

2.D. Experimental setup

2.D.1. Evaluation method

In this section, we present the way we evaluate our CAD systems, focusing on detection criteria. We avoid using a voxel-based ROC or simple Jaccard measures or Dice coefficient scores due to the fact that otherwise the results would be biased toward larger WMHs, since these contain more voxels. Instead we adapt an FROC analysis to assess the system detection performance. The following details how we calculate the FROC: We first create candidate segments by accepting voxels with likelihoods higher than

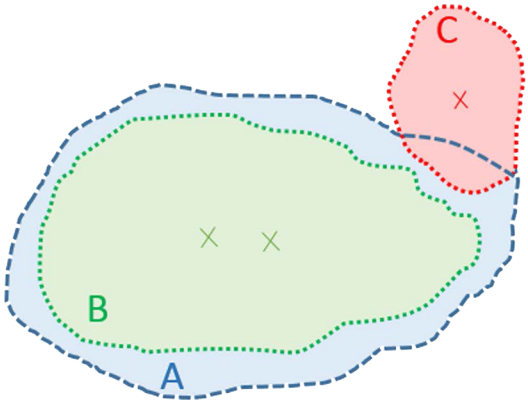


FIG. 6. An abstract example depicting a WMH segment (A) in the reference annotation, and two corresponding candidate segments (B) and (C). The crosses show the segments' representative voxels. Reference standard segment (A) is considered as a true positive, since it is hit by some of the candidate segments' representative voxels. Unlike (B), (C) is counted as a false positive since at least one of its representative voxels is out of the reference standard WMH annotation.

a threshold t in the likelihood map, which is the soft classification result on each test subject for the classifier to be evaluated. Then each resulting candidate segment is assigned the likelihood of the most likely WMH voxel inside that candidate. At a given analysis threshold t', we remove all of the candidate segments that are assigned likelihoods smaller than t' and subsequently we calculate true positive rate and average number of false positives per patient as follows: We select inside each candidate segment the voxels that are the local maxima of Euclidean distance of each voxel to the boundary of the candidate. Then these representative voxels are investigated to determine if they are marked as WMH in the reference standard or not. If any of them is not marked as WMH, we consider the candidate as a false positive. WMH segments in the reference standard that are not detected by any representative voxels of the candidate segments are considered to be false negatives. Figure 6 illustrates an example for a better understanding of this procedure.

The FROC curve is obtained by varying the analysis threshold t' between 0 and 1. Notice that the threshold t to create candidate segments from the likelihood map is kept constant during the analysis, and is different from the analysis threshold t', which varies to generate the curve. In order to suppress the effect of t across different methods, we fix t such that the total volume of all created segments is as close as possible to the total volume of WMHs in the reference standard.

We compute p-values for statistical significance tests as follows: We create 100 bootstraps by sampling subjects on the test set with replacements. Then the area under the FROC curves was computed on each bootstrap for each of the two compared methods. Empirical p-values were reported as the proportion of bootstraps where the area under the FROC curve for method B was higher than A, when the null-hypothesis to reject was "method A is no better than B." If no such bootstrap existed, the p-value <0.01 was reported, representing a significant difference.

2.D.2. Comparisons

We evaluate the performance of the proposed method using the FROC analysis, as introduced in Subsection 2.D.1 and compare its performance to a number of surrogates. Most importantly as two human reader annotations are available on the test set, we compare the performance of the method to the human readers. We also evaluate the effect of Adaboost classifier used in the method and compare its performance to the cases where a single random forest or a Gentleboost classifier with 100 decision stumps as the basic classifier are trained. We also compare the results with W2MHS, a recent publicly available automatic lesion segmentation package,[28] which applies a random forest (with 50 subtrees, $\sqrt{\#features}$ features randomly selected at each node and information gain tree splitting criterion) on texture and intensity-variation based features.

Each of the mentioned comparisons is made separately for each size category and all of the lesions and twice considering reader 1 and reader 2 as the reference standard, together with the average of the two cases.

To assess the robustness of the algorithm for cases with motion artifacts, noise, or failure at one of the preprocessing steps, the algorithm was evaluated both on cases with and without these artifacts (see subsection 2.B.3), and we also present a comparison to performance of the independent human reader.

As a strategy of our methodology, we train a two-stage classifier. To assess the effectiveness of this method ingredient, we also train a single-stage classifier on the whole dataset with the same feature set and the same type of classifier (five iterations of Adaboost on random forests) and compare the results.

3. RESULTS

Figures 7(a)–7(c) present the FROC curves with 95% confidence intervals for detection of large WMHs and compare the performance of the proposed method to the performance of human readers, two other classifiers (random forest and Gentleboost using 100 regression stumps as its weak classifiers), and the W2MHS method.[28] The same experiments were repeated for detection of small WMHs as presented in Figs. 7(d)–7(f). Figures 7(g)–7(i) represent the same for detection of all WMHs with the second-stage classifier.

An FROC comparison for the performance of the system on normal and harder cases is depicted in Fig. 8. Figure 9 investigates the effect of the size-based separation strategy used in our research on detection of all of the WMHs [Fig. 9(a)] and detection of small WMHs [Fig. 9(b)]. The differences are statistically significant in both cases (p-value < 0.01). In Fig. 10, a number of sample FLAIR slices from three of the patients, together with the detections of the CAD system and the annotations of the two human readers, are shown for a qualitative comparison.

After system evaluation, we had a closer look at the false positives of the system. Observing the false positives showed that in a considerable proportion of cases, the underlying tissue was suspicious. Based on this, we asked an expert neurologist

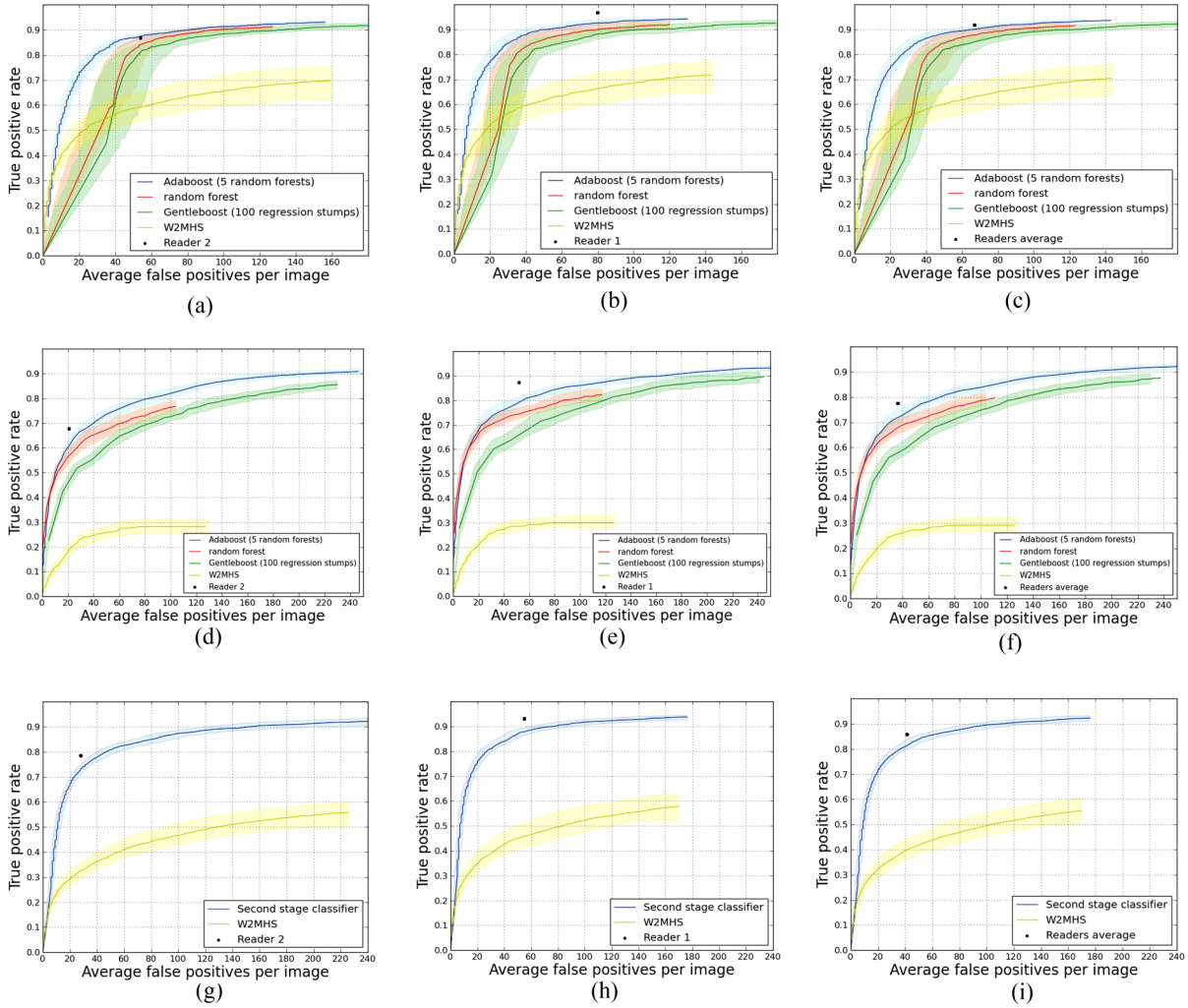


FIG. 7. FROC curves with 95% confidence intervals that compare the performance of different classifiers and human readers on detection of large [(a)–(c)], small [(d)–(f)] and all of the WMHs [(g)–(i)]. First and second columns are evaluated considering reader 1 and reader 2 as the reference standard. The third column represents the average.

to either accept or reject false positives as true WMHs on all of test cases. As a result, on average 15.1 false positives per patient and in a subject more than 50 false positives were accepted.

To show the size specific performance of CAD system, we performed a size-based analysis of TP rate, which is depicted in Fig. 11.

4. DISCUSSION

4.A. Data acquisition matters

In order to train and evaluate our algorithm, we made use of a dataset containing 362 MRI scans of SVD patients. Use of hundreds of subjects for the development of these algorithms is not seen in other studies of WMH detection.

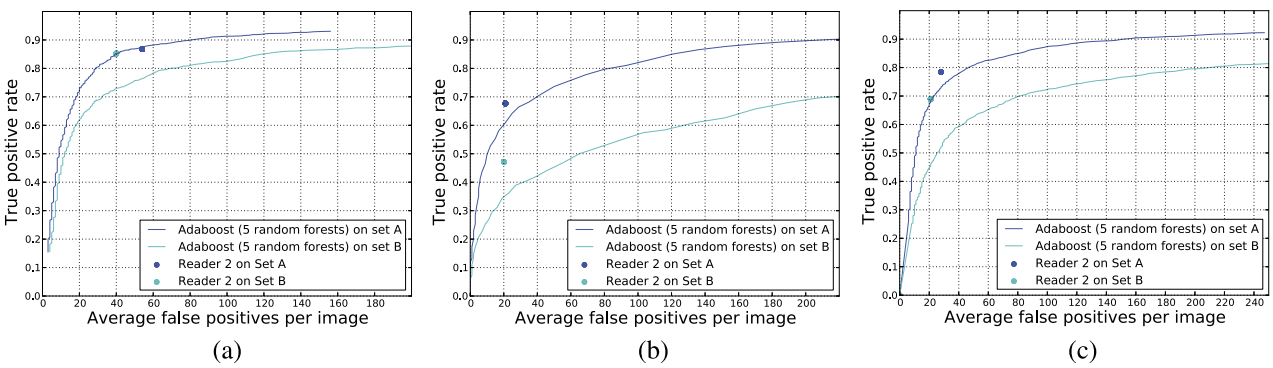


FIG. 8. FROC curves representing performance of the proposed CAD system for normal cases (set A) compared to the set of harder cases (set B), with reader 1 annotations as the reference standard. (a) Detection of large WMHs, (b) detection of small WMHs, and (c) detection of all WMHs.

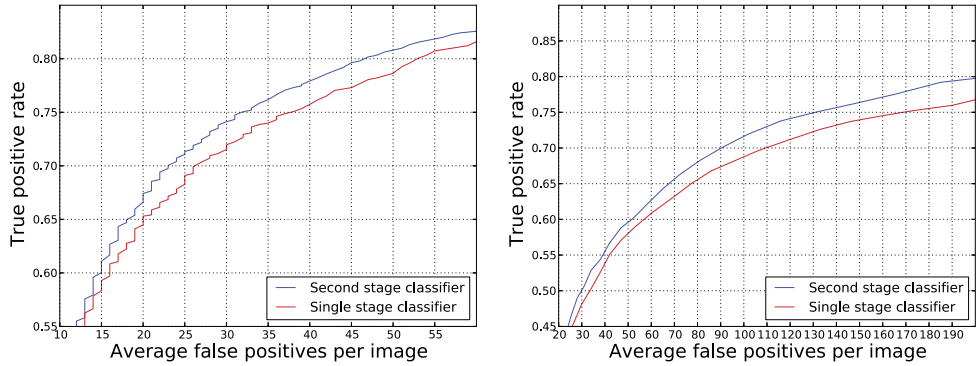


FIG. 9. FROC curves that compare the performance of the combined small- and large lesions classification results versus a single stage classifiers for detection of all and small WMHs (smaller than 3 mm in in-plane effective diameter), considering reader 1 as the reference standard. (a) Detection of all WMHs and (b) small WMH detection.

This large dataset has aided in better generalization and made it possible to avoid overfitting of the model to the noise patterns. On the other hand the acquisitions used in this study were made in 2006 on a 1.5 T MR machine and the FLAIR acquisitions in particular have a relatively high slice thickness of 5 mm with 1 mm of interslice gap. More modern acquisition protocols together with higher field strength MR systems lead to a smaller slice thickness. This reduces the partial volume effect observed in smaller WMHs.

Our algorithm has been developed to work slice based because of the thicker FLAIR slices. Iso-voxel, fine resolution, FLAIR scans enable the use of 3D features. The same methodology can be used with updated features to fully benefit from these 3D acquisitions.

Furthermore, a more accurate ground truth, especially on smaller WMHs, could have helped a more accurate evaluation. Such improvements on the ground truth can be achieved using a consensus of readers or including more readers, though this might be expensive on large datasets.

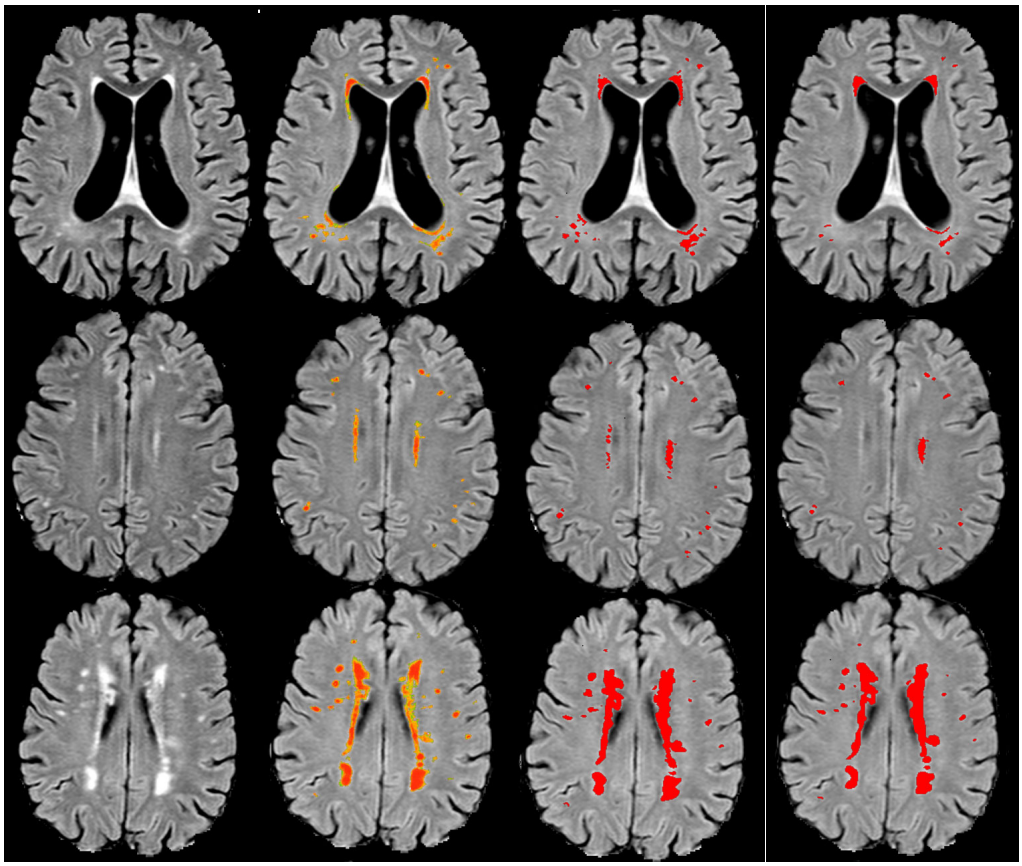


FIG. 10. A demonstration of our CAD system detection together with human readers annotations. (a) FLAIR images without annotations, (b) likelihood maps provided by CAD, (c) annotations by human reader 1, and (d) annotations by human reader 2.

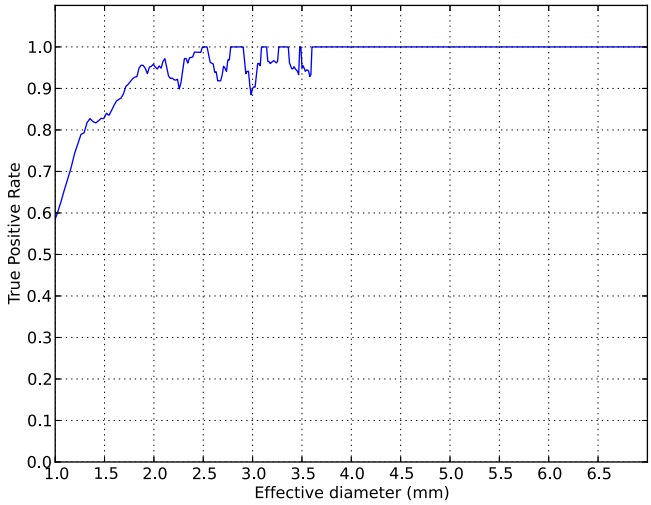


FIG. 11. True positive rate for different WMH sizes (at a total true positive rate of 0.75).

4.B. Single or two stage classification?

There were two method ingredients in our approach that resulted in a competitive performance for the single stage classifier: First the set of features used for the single classifier includes features optimized for detection of small WMHs, and second the usage of Adaboost on top of random forests emphasizes the detection of harder samples. Even though our results show that the single stage classifier can be considered as a reliable option, the two-stage classification scheme results in a better detection of small WMHs. Considering the true positive rates in range of 0.75–0.85, which seems reasonable to be used in practice, the two-stage classification scheme on average results in 13% less false positives for every detection point in the same true positive rates in the mentioned range, though they perform similarly for TP rates below 50%. Also p-value of <0.01 showed a statistically significant improvement.

Noting the heterogeneity of the appearances of smaller and larger WMHs, the representation of lesions would be scattered over the feature space resulting in a highly nonlinear problem for a single classifier. By training specified classifiers for the two heterogeneous categories and training the second-stage classifier given the likelihoods of each category, the nonlinearity of the problem is reduced on the new feature space and therefore we expect the two-stage classification scheme to result in an improvement, as observed empirically by the results presented in Fig. 9.

4.C. Accurate detection of smaller lesions is more challenging

As the comparison of the detection curves for small and large WMHs in Fig. 7 suggests, detection of small WMHs is a much more complicated task for which we hypothesize the following reasons: First of all the partial volume effect causes small WMHs to appear in less contrast to normal white matter. Second, noise in the image might appear similar to small WMHs. And finally, small WMHs are much more prone to be missed by the human readers compared to large WMHs (see Fig. 1). This results in an inconsistent training dataset where some true small WMHs are labeled as negative samples, which might be confusing for the classifier.

4.D. Comparison to other methods

A multitude of automated detection algorithms for WMHs exists, but since most of the current automated approaches are tuned to optimize segmentation performance according to Jaccard or Dice scores, smaller WMHs often go undetected in these approaches.

Generalized test datasets to compare performance of different WMH segmentation/detection algorithms do not exist for diseases other than multiple sclerosis. Lesions in MS are mostly different in their size, appearance, and localization from lesions that are seen in SVD. Therefore it is not desirable to use existing test databases (such as the MICCAI MS lesion segmentation challenge 2008 or the ISBI 2015 Longitudinal MS Lesion Segmentation Challenge), nor would it be fair to expect compatible results from algorithms designed for lesions caused by different underlying pathology. To provide some results, we compare the performance of our algorithm with the publicly available W2MHS algorithm (Fig. 7).

4.E. On potential importance of smaller lesions

Several important ingredients of the proposed method are optimized for an accurate detection of all-size lesions, large ones as well as small ones. The main importance of detecting these small WMHs is their etiological importance. By detecting these small lesions it is possible to follow WMH growth and progression in follow-up studies, even per location, and with that gain more knowledge about the development of WMHs. It might be the case that intervening at a relatively early stage could prevent progression of small WMHs, possibly averting progression in clinical symptoms. This mechanism is still speculative and needs further investigation. These future investigations rely on the accurate detection and localization of small WMHs, as presented in this paper.

5. CONCLUSIONS

In this paper, a fully automated system for detection of WMHs was presented that uses a two-stage classification approach, based on combining two size-specific classifiers. Experiments show that the proposed CAD system performs close to human readers. Ingredients of the method were chosen to enable the CAD system to accurately detect small WMHs as well as the larger ones. This includes the set of features, classifier type, and the two-stage classification scheme based on small and large WMH detectors.

The effect of these factors was investigated and shown to be contributing to better detection of WMHs. Our system reaches a true positive rate of 0.80 with 47 and 27 false positives per volume using reader 1 and reader 2 as the reference standard, respectively. The real performance of the classifier could

be potentially better if a more accurate reference standard, especially on detection of small WMHs, was available.

ACKNOWLEDGMENTS

This work was supported by a VIDI innovational grant from the Netherlands Organisation for Scientific Research (NWO, Grant No. 016.126.351). The authors also would like to acknowledge Lucas J.B. van Oudheusden and Renate M. Arntz for their contributions to this study.

CONFLICT OF INTEREST DISCLOSURE

The authors have no COI to report.

[a]Electronic mail: mohsen.ghafoorian@radboudumc.nl

[1]F. De Leeuw, J. C. de Groot, E. Achten, M. Oudkerk, L. Ramos, R. Heijboer, A. Hofman, J. Jolles, J. Van Gijn, and M. Breteler, "Prevalence of cerebral white matter lesions in elderly people: A population based magnetic resonance imaging study. The rotterdam scan study," J. Neurol., Neurosurg. Psychiatry **70**(1), 9–14 (2001).

[2]J. M. Wardlaw, E. E. Smith, G. J. Biessels, C. Cordonnier, F. Fazekas, R. Frayne, R. I. Lindley, J. T. O'Brien, F. Barkhof, O. R. Benavente, S. E. Black, C. Brayne, M. Breteler, H. Chabriat, C. Decarli, F. E. de Leeuw, F. Doubal, M. Duering, N. C. Fox, S. Greenberg, V. Hachinski, I. Kilimann, V. Mok, R. v. Oostenbrugge, L. Pantoni, O. Speck, B. C. Stephan, S. Teipel, A. Viswanathan, D. Werring, C. Chen, C. Smith, M. van Buchem, B. Norrving, P. B. Gorelick, and M. Dichgans, "Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration," Lancet Neurol. **12**(8), 822–838 (2013).

[3]H. Baezner, C. Blahak, A. Poggesi, L. Pantoni, D. Inzitari, H. Chabriat, T. Erkinjuntti, F. Fazekas, J. Ferro, P. Langhorne, J. O'Brien, P. Scheltens, M. Visser, L. Wahlund, G. Waldemar, A. Wallin, and M. Hennerici, "Association of gait and balance disorders with age-related white matter changes the ladis study," Neurology **70**(12), 935–942 (2008).

[4]J. C. de Groot, M. Oudkerk, J. v. Gijn, A. Hofman, J. Jolles, and M. Breteler, "Cerebral white matter lesions and cognitive function: The rotterdam scan study," Ann. Neurol. **47**(2), 145–151 (2000).

[5]I. W. van Uden, A. M. Tuladhar, K. F. de Laat, A. G. van Norden, D. G. Norris, E. J. van Dijk, I. Tendolkar, and F.-E. de Leeuw, "White matter integrity and depressive symptoms in cerebral small vessel disease: The run DMC study," Am. J. Geriatr. Psychiatry **23**(5), 525–535 (2015).

[6]M. van Zagten, J. Lodder, and F. Kessels, "Gait disorder and parkinsonian signs in patients with stroke related to small deep infarcts and white matter lesions," Mov. Disord. **13**(1), 89–95 (1998).

[7]S. E. Vermeer, N. D. Prins, T. den Heijer, A. Hofman, P. J. Koudstaal, and M. M. Breteler, "Silent brain infarcts and the risk of dementia and cognitive decline," N. Engl. J. Med. **348**(13), 1215–1222 (2003).

[8]L. Pantoni, A. M. Basile, G. Pracucci, K. Asplund, J. Bogousslavsky, H. Chabriat, T. Erkinjuntti, F. Fazekas, J. M. Ferro, M. Hennerici, J. O'Brien, P. Scheltens, M. Visser, L.-O. Wahlund, G. Waldemar, A. Wallin, and D. Inzitari, "Impact of age-related cerebral white matter changes on the transition to disability—The LADIS study: Rationale, design and methodology," Neuroepidemiology **24**(1-2), 51–62 (2004).

[9]A. G. van Norden, K. F. de Laat, R. A. Gons, I. W. van Uden, E. J. van Dijk, L. J. van Oudheusden, R. A. Esselink, B. R. Bloem, B. G. van Engelen, M. J. Zwarts, I. Tendolkar, M. G. Olde-Rikkert, M. J. van der Vlugt, M. P. Zwiers, D. G. Norris, and F. E. de Leeuw, "Causes and consequences of cerebral small vessel disease. The run DMC study: A prospective cohort study. Study rationale and protocol," BMC Neurol. **11**(1), 29–39 (2011).

[10]M. M. Schoonheim, R. M. Vigeveno, F. C. R. Lopes, P. J. Pouwels, C. H. Polman, F. Barkhof, and J. J. Geurts, "Sex-specific extent and severity of white matter damage in multiple sclerosis: Implications for cognitive decline," Hum. Brain Mapp. **35**(5), 2348–2358 (2014).

[11]N. Hirono, H. Kitagaki, H. Kazui, M. Hashimoto, and E. Mori, "Impact of white matter changes on clinical manifestation of Alzheimer's disease a quantitative study," Stroke **31**(9), 2182–2188 (2000).

[12]C. D. Smith, D. A. Snowdon, H. Wang, and W. R. Markesbery, "White matter volumes and periventricular white matter hyperintensities in aging and dementia," Neurology **54**(4), 838–842 (2000).

[13]G. Weinstein, A. S. Beiser, C. DeCarli, R. Au, P. A. Wolf, and S. Seshadri, "Brain imaging and cognitive predictors of stroke and Alzheimer disease in the framingham heart study," Stroke **44**(10), 2787–2794 (2013).

[14]G. Marshall, E. Shchelchkov, D. Kaufer, L. Ivanco, and N. Bohnen, "White matter hyperintensities and cortical acetylcholinesterase activity in parkinsonian dementia," Acta Neurol. Scand. **113**(2), 87–91 (2006).

[15]G. Whitman, T. Tang, A. Lin, and R. Baloh, "A prospective study of cerebral white matter abnormalities in older people with gait dysfunction," Neurology **57**(6), 990–994 (2001).

[16]L. L. Herrmann, M. Le Masurier, and K. P. Ebmeier, "White matter hyperintensities in late life depression: A systematic review," J. Neurol., Neurosurg. Psychiatry **79**(6), 619–624 (2008).

[17]F. Admiraal-Behloul, D. Van Den Heuvel, H. Olofsen, M. J. van Osch, J. van der Grond, M. Van Buchem, and J. Reiber, "Fully automatic segmentation of white matter hyperintensities in MR images of the elderly," NeuroImage **28**(3), 607–617 (2005).

[18]S. Jain, D. M. Sima, A. Ribbens, M. Cambron, A. Maertens, W. Van Hecke, J. De Mey, F. Barkhof, M. D. Steenwijk, M. Daams, F. Maes, S. Van Huffel, H. Vrenken, and D. Smeets, "Automatic segmentation and volumetry of multiple sclerosis brain lesions from MR images," NeuroImage: Clin. **8**, 367–375 (2015).

[19]A. Khademi, A. Venetsanopoulos, and A. R. Moody, "Robust white matter lesion segmentation in flair MRI," IEEE Trans. Biomed. Eng. **59**(3), 860–871 (2012).

[20]L. Shi, D. Wang, S. Liu, Y. Pu, Y. Wang, W. C. Chu, A. T. Ahuja, and Y. Wang, "Automated quantification of white matter lesion in magnetic resonance imaging of patients with acute infarction," J. Neurosci. Methods **213**(1), 138–146 (2013).

[21]N. Shiee, P.-L. Bazin, A. Ozturk, D. S. Reich, P. A. Calabresi, and D. L. Pham, "A topology-preserving approach to the segmentation of brain images with multiple sclerosis lesions," NeuroImage **49**(2), 1524–1535 (2010).

[22]K. Van Leemput, F. Maes, D. Vandermeulen, A. Colchester, and P. Suetens, "Automated segmentation of multiple sclerosis lesions by model outlier detection," IEEE Trans. Med. Imaging **20**(8), 677–688 (2001).

[23]R. Khayati, M. Vafadust, F. Towhidkhah, and M. Nabavi, "Fully automatic segmentation of multiple sclerosis lesions in brain MR flair images using adaptive mixtures method and markov random field model," Comput. Biol. Med. **38**(3), 379–390 (2008).

[24]R. de Boer, H. A. Vrooman, F. van der Lijn, M. W. Vernooij, M. A. Ikram, A. van der Lugt, M. Breteler, and W. J. Niessen, "White matter lesion extension to automatic brain tissue segmentation on MRI," NeuroImage **45**(4), 1151–1161 (2009).

[25]P. Schmidt, C. Gaser, M. Arsic, D. Buck, A. Förschler, A. Berthele, M. Hoshi, R. Ilg, V. J. Schmid, C. Zimmer, B. Hemmer, and M. Mhlau, "An automated tool for detection of flair-hyperintense white-matter lesions in multiple sclerosis," NeuroImage **59**(4), 3774–3783 (2012).

[26]J.-Z. Tsai, S.-J. Peng, Y.-W. Chen, K.-W. Wang, C.-H. Li, J.-Y. Wang, C.-J. Chen, H.-J. Lin, E. E. Smith, H.-K. Wu, S.-F. Sung, P.-S. Yeh, and Y.-L. Hsin, "Automated segmentation and quantification of white matter hyperintensities in acute ischemic stroke patients with cerebral infarction," PloS One **9**(8), e104011 (2014).

[27]S. Klöppel, A. Abdulkadir, S. Hadjidemetriou, S. Issleib, L. Frings, T. N. Thanh, I. Mader, S. J. Teipel, M. Hüll, and O. Ronneberger, "A comparison of different automated methods for the detection of white matter lesions in MRI data," NeuroImage **57**(2), 416–422 (2011).

[28]V. Ithapu, V. Singh, C. Lindner, B. P. Austin, C. Hinrichs, C. M. Carlsson, B. B. Bendlin, and S. C. Johnson, "Extracting and summarizing white matter hyperintensities using supervised segmentation methods in Alzheimer's disease risk and aging studies," Hum. Brain Mapp. **35**(8), 4219–4235 (2014).

[29]M. M. Riad, B. Platel, F.-E. de Leeuw, and N. Karssemeijer, "Detection of white matter lesions in cerebral small vessel disease," Proc. SPIE **8670**, 867014 (2013).

[30]A. P. Zijdenbos and B. M. Dawant, "Brain segmentation and white matter lesion detection in MR images," Crit. Rev. Biomed. Eng. **22**(5-6), 401–465 (1993).

[31]Z. Karimaghaloo, M. Shah, S. J. Francis, D. L. Arnold, D. L. Collins, and T. Arbel, "Automatic detection of gadolinium-enhancing multiple sclerosis lesions in brain MRI using conditional random fields," IEEE Trans. Med. Imaging **31**(6), 1181–1194 (2012).

[32]E. Geremia, B. H. Menze, O. Clatz, E. Konukoglu, A. Criminisi, and N. Ayache, "Spatial decision forests for MS lesion segmentation in multi-channel MR images," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2010* (Springer, Berlin Heidelberg, 2010), pp. 111–118.
[33]P. Anbeek, K. L. Vincken, M. J. van Osch, R. H. Bisschops, and J. van der Grond, "Probabilistic segmentation of white matter lesions in MR imaging," NeuroImage **21**(3), 1037–1044 (2004).
[34]M. D. Steenwijk, P. J. Pouwels, M. Daams, J. W. van Dalen, M. W. Caan, E. Richard, F. Barkhof, and H. Vrenken, "Accurate white matter lesion segmentation by k nearest neighbor classification with tissue type priors (kNN-TTPs)," NeuroImage: Clin. **3**, 462–469 (2013).
[35]Z. Karimaghaloo, H. Rivaz, D. L. Arnold, D. L. Collins, and T. Arbel, "Temporal hierarchical adaptive texture crf for automatic detection of gadolinium-enhancing multiple sclerosis lesions in brain MRI," IEEE Trans. Med. Imaging **34**(6), 1227–1241 (2015).
[36]M. E. Caligiuri, P. Perrotta, A. Augimeri, F. Rocca, A. Quattrone, and A. Cherubini, "Automatic detection of white matter hyperintensities in healthy aging and pathology using magnetic resonance imaging: A review," Neuroinformatics **13**(3), 1–16 (2015).
[37]D. García-Lorenzo, S. Francis, S. Narayanan, D. L. Arnold, and D. L. Collins, "Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging," Med. Image Anal. **17**(1), 1–18 (2013).
[38]R. Schmidt, P. Scheltens, T. Erkinjuntti, L. Pantoni, H. Markus, A. Wallin, F. Barkhof, and F. Fazekas, "White matter lesion progression a surrogate endpoint for trials in cerebral small-vessel disease," Neurology **63**(1), 139–144 (2004).
[39]E. J. Hoffman, S.-C. Huang, and M. E. Phelps, "Quantitation in positron emission computed tomography: 1. Effect of object size," J. Comput. Assisted Tomogr. **3**(3), 299–308 (1979).
[40]J. E. Moody, "Note on generalization, regularization and architecture selection in nonlinear learning systems," in *Proceedings of IEEE Workshop Neural Networks for Signal Processing [1991]* (IEEE, Princeton, New Jersey, 1991), pp. 1–10.
[41]P. C. Bunch, J. F. Hamilton, G. K. Sanderson, and A. H. Simmons, "A free response approach to the measurement and characterization of radiographic observer performance," Proc. SPIE **0127**, 124–135 (1977).
[42]K. W. Kim, J. R. MacFall, and M. E. Payne, "Classification of white matter lesions on magnetic resonance imaging in elderly persons," Biol. Psychiatry **64**(4), 273–280 (2008).
[43]M. Jenkinson and S. Smith, "A global optimisation method for robust affine registration of brain images," Med. Image Anal. **5**(2), 143–156 (2001).
[44]J. Mazziotta, A. Toga, A. Evans, P. Fox, J. Lancaster, K. Zilles, R. Woods, T. Paus, G. Simpson, B. Pike, C. Holmes, L. Collins, P. Thompson, D. MacDonald, M. Iacoboni, T. Schormann, K. Amunts, N. Palomero-Gallagher, S. Geyer, L. Parsons, K. Narr, N. Kabani, G. Le Goualher, J. Feidler, K. Smith, D. Boomsma, H. H. Pol, T. Cannon, R. Kawashima, and B. Mazoyer, "A four-dimensional probabilistic atlas of the human brain," J. Am. Med. Inf. Associat. **8**(5), 401–430 (2001).
[45]S. M. Smith, "Fast robust automated brain extraction," Hum. Brain Mapp. **17**(3), 143–155 (2002).
[46]Y. Zhang, M. Brady, and S. Smith, "Segmentation of brain MR images through a hidden markov random field model and the expectation–maximization algorithm," IEEE Trans. Med. Imaging **20**(1), 45–57 (2001).
[47]G. J. McLachlan and K. E. Basford, *Mixture Models. Inference and Applications to Clustering, Statistics: Textbooks and Monographs* (Dekker, New York, 1988), p. 1.
[48]M. Ghafoorian, N. Karssemeijer, I. van Uden, F. E. de Leeuw, T. Heskes, E. Marchiori, and B. Platel, "Small white matter lesion detection in cerebral small vessel disease," Proc. SPIE **9414**, 941411 (2015).
[49]R. Moshavegh, B. Bejnordi, A. Mehnert, K. Sujathan, P. Malm, and E. Bengtsson, "Automated segmentation of free-lying cell nuclei in pap smears for malignancy-associated change analysis," in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (IEEE, San Diego, 2012), pp. 5372–5375.
[50]A. Kuijper, "Geometrical pdes based on second-order derivatives of gauge coordinates in image processing," Image Vis. Comput. **27**(8), 1023–1034 (2009).
[51]L. Breiman, "Random forests," Mach. Learn. **45**(1), 5–32 (2001).
[52]Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," J. Comput. Syst. Sci. **55**(1), 119–139 (1997).
[53]J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: A statistical view of boosting (with discussion and a rejoinder by the authors)," Ann. Stat. **28**(2), 337–407 (2000).
[54]M. Ghafoorian, N. Karssemeijer, T. Heskes, I. W. M. van Uder, F. E. de Leeuw, E. Marchiori, B. van Ginneken, and B. Platel, "Non-uniform patch sampling with deep convolutional neural networks for white matter hyperintensity segmentation," in *Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on* (IEEE, 2016), pp. 1414–1417.
[55]M. Ghafoorian, N. Karssemeijer, T. Heskes, I. van Uden, C. Sanchez, G. Litjens, F.-E. de Leeuw, B. van Ginneken, E. Marchiori, and B. Platel, "Location Sensitive Deep Convolutional Neural Networks for Segmentation of White Matter Hyperintensities," preprint arXiv:1610.04834 (2016).